DOCUMENT RESUME

ED 208 036 TM 810 733

AUTHOR Bode, Rita K.
TITLE Comparison of Pretest and Reanalysis Results of an Item Bias Study.
PUB DATE Apr 81
NOTE 31p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; Intermediate Grades; *Item Analysis; *Racial Bias; *Sex Bias; *Test Bias; Test Items
IDENTIFIERS *SRA Achievement Series

ABSTRACT

This presentation compares the results of a reanalysis of items in one form of a level of the 1978 SRA Achievement Series with the results from the pretest analysis. The bias analyses were performed on white or majority, black, Hispanic, male, and female samples from the pretest and two subsequent samples. Comparisons were made in terms of the consistency of identification of biased items, the direction of bias, and baseline comparisons across samples within a sex or racial/ethnic group. Results show less than expected consistency of identification; effective balancing of items in direction of bias; the importance of sample size in consistency; and inconclusive results in the baseline comparisons. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED208036

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Comparison of Pretest and Reanalysis Results of an Item Bias Study

Rita K. Bode

Science Research Associates, Inc.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
R. K. Bode
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Presented at the annual meeting of the American Educational Research Association, Los Angeles, April 1981.

TM 810 733

Comparison of Pretest and Reanalysis Results of an Item Bias Study

Bias studies are now in the stage where practical considerations are being investigated. Recent investigators of statistical bias techniques have concentrated on comparisons of various methods to determine bias on the same sets of data--either real data (Shepard, Camilli & Averill, 1980; Ironson & Subkoviak, 1979; Rudner & Convey, 1978) or data simulated using the Monte Carlo method (Merz & Grossen, 1979; Rudner, Getson & Knight, 1980). Few studies (Scheuneman, 1980) have investigated the same technique of determining bias on various sets of data although Angoff (1980) discusses such comparisons.

In test development the time frame is usually short so once an item is identified as "biased", it is eliminated from the item pool. In the development of achievement tests, however, there are exceptions: when the item measures content that is considered necessary for complete subject area coverage or for skill areas that are important to measure. In such cases, items that are biased in the opposite direction are included in the final test forms to counterbalance the presence of the first set of biased items.

The procedures just described are based on the assumption that the same results would be obtained if the bias study were

conducted on comparable data samples. In this study, the assumption of comparable results using subsequent samples is being tested. Items that were analyzed for biasas part of the pretest and item selection for form 1 of Level E (appropriate for grades 4-5) of the 1978 edition of the SRA Achievement Series were reanalyzed on two subsequent samples (subsamples of the spring and fall standardization samples). For the pretest and each reanalysis, data were available for samples of majority, white, black, Hispanic, male, and female samples.

Data from majority/minority samples (white-black, white-Hispanic, and male-female) were compared across the three sets of data to determine the degree of consistency of identification of "biased" items. (Note: In this study, the female sample is being considered the minority sample.) In addition to these comparisons, data from the fall majority/spring majority and fall minority/spring minority samples were also compared to investigate variation due to sampling (Angoff, 1980). Also investigated in the comparison groups were differences in distractor selection, item difficulty, item discrimination, percentage of omits, and percentage of "not reached," and possible patterns in skill areas measured by the items identified as "biased".

Pretest Analysis

As part of the selection of items for the final test forms, statistical studies were conducted to identify and eliminate items that were identified as biased. (The content or editorial bias analysis that was conducted on all pretest items is not discussed here.) In this analysis, the definition of bias that was used is as follows: "an item of a test is said to be biased for members of a particular group if, on that item, the members of the group obtain an average score which differs from the average score of the other groups by more or less than expected from performance on other items of the same test." (Cleary, 1968).

Samples for the sex and racial/ethnic bias studies were drawn from the 1977 pretest sample. The pretest sample consisted of 48,524 students from 979 classes in 226 schools from 154 cities in 38 states. Schools selected for the majority and minority samples represented their geographic areas in size, location, ethnic distribution, and student ability.

The item bias study for Level E included 3505 students in the majority sample; 664 students in the black sample, 522 students in the Hispanic sample, and 101 students in the American Indian sample (data from the American Indian sample were excluded from this study because of insufficient sample size). The majority samples consisted of students in schools that were predominantly

white or racially/ethnically mixed, with no one group predominant. The minority samples consisted of students from schools whose populations were predominantly black or Hispanic. Male and female samples were identified from coding on the individual answer sheets of students in the majority sample only.

Pretest units were randomly assigned to the majority and minority samples. Each pretest unit was not necessarily administered to each minority sample but it was given to at least one minoirity group sample. The pretest bias study combined items from all pretest units into one set of items for each comparison. Because not all minority groups took all units, the set of items for the majority/black comparison was different from the set of items used for the majority/Hispanic comparison. Only the male/female comparison contained the complete set of items. As a consequence, twelve of the items identified as biased in subsequent analyses were found not to have been pretested with the minority group for which they were biased.

Items analyzed had survived a previous screening that met the following statistical criteria: $\underline{p}$ values between .30 and .80, point biserial item-test correlations greater than .30, and a positive $\underline{p}$ value gain from one grade to the next. Items that did not meet these criteria would not be included in the final test forms and were thus eliminated before the bias studies were conducted.

For the items that met the statistical criteria, *p* values for the male, female, majority, and minority samples were converted to normalized *z* scores with a mean of 0 and a standard deviation of 1. The *z* scores were plotted separately for male versus female samples, majority versus black samples, and majority versus Hispanic samples.

For each comparison made, an ellipse was drawn around the clustering of item plots along the apparent best-fit line. The ellipse was drawn as narrowly as possible while including as many items as possible. The actual shape and width of the ellipse was based on an inspection of the clustering of item plots. The outliers were defined as biased items. The items that were close to the perimeter of the ellipse were judged marginal.

The disposition of items in the Level E pretest bias analysis for the male-female, majority-black, and majority-Hispanic samples is presented in Table 1. The data consist of the number of items analyzed, the number identified as biased, and the number of biased items used in the final forms.

TABLE 1

Item Disposition from 1977 Achievement Series Pretest

Level E Subtests	Male-Female			Majority-Black			Majority-Hispanic			Total	
	A[a]	B	U	A	B	U	A	B	U	B	U
Reading Vocabulary	136	1	0	116	9	0	102	23	0	29	0
Reading Comprehension	220	3	0	191	14	1	161	12	2	23	3
Math Concepts	38	3	0	38	5	0	38	5	0	12	0
Math Computation	114	6	1	96	8	0	82	13	0	23	1
Math Problem Solving	151	12	1	133	8	0	111	8	1	22	2
Lang. Arts Mechanics	175	2	0	154	21	0	133	16	1	31	1
Lang. Arts Usage	148	6	0	130	17	0	111	8	0	26	0
Lang. Arts Spelling	125	16	0	103	9	0	125	26	2	63	2
Reference Materials	279	16	1	237	19	0	198	24	0	54	1
Social Studies	174	7	1	146	13	4	107	10	0	23	5
Science	153	4	3	127	15	4	105	13	4	26	10

[a]A = Analyzed; B = Biased; U = Used

Spring and Fall Reanalysis

After the standardization of the test, the bias studies were repeated using a variation of the delta-plot method which consists of formal procedures for measuring departure of items from the plot (Angoff and Ford, 1973). The purpose of the reanalysis was to determine the effectiveness of the pretest analysis in identifying and eliminating biased items from the final test forms. Two additional samples were used in the reanalysis because of inconsistencies in sample definition between the pretest and reanalysis. The samples for the three analyses were different in several ways. The samples differed in time: pretest was in April 1977; spring standardization was in April 1978; and fall standardization was in October 1978. Pretest minority samples were identified on the basis of school demographics, that is, if the majority of students in the school were black, the data for all students in that school were included in the black sample. In the reanalysis samples, the minority samples consisted only of minority students because racial/ethnic and sex information was collected for individual students. Lastly, the female and male samples in the spring and fall analyses were sampled across racial/ethnic groups.

The reanalysis was performed on subsamples of form 1 data from the spring and fall standardization samples. The spring standardization sample consisted of 83,681 students (3713 taking Level E in grade 4) in 383 schools in 81 districts from 36

states; the fall standardization sample consisted of 129,900 students (6800 taking Level E in grade 5) in 457 schools in 92 districts from 38 states. The spring and fall reanalysis samples were each selected from their respective national standardization samples in such a way that their score distribution matched that of the complete standardization sample. Student data were sorted on a random variable and the students pulled to meet the distributional requirements. Majority and minority samples, as well as male and female samples, were identified from coding on student answer sheets.

Item $\bar{p}$ values for each sample were converted to delta values (normalized standard scores with a mean of 13 and a standard deviation of 4). Data were analyzed separately for the male versus female samples, white versus black samples, and white versus Hispanic samples. In addition to plotting the delta values for each comparison group, the major axis was computed and the perpendicular distance (D) between the major axis and item plots was calculated. The calculated D values were used in all subsequent analyses and summarizations of results.

In the reanalysis, an item was identified as biased if the D value was equal to or greater than plus or minus 1.00. Based on this definition, 12 items were identified as biased in the spring analysis and 20 in the fall analysis. The number of biased items, by test and comparison, are presented in Table 2.

TABLE 2

Number of Biased Items in Reanalysis[b]

	Comparison Groups					
	B/W		F/M		H/W	
Test	S	F	S	F	S	F
Vocabulary	2	2	0	0	0	3
Comprehension	1	0	0	0	0	2
Math Concepts	1	2	0	0	0	0
Computation	0	0	0	0	2	4
Problem Solving	0	0	0	0	0	0
Mechanics	0	0	0	0	0	1
Usage	3	1	0	0	0	0
Spelling	0	0	0	0	0	1
Reference Mtls.	1	1	0	0	0	1
Social Studies	2	0	0	0	0	1
Science	0	0	0	1	0	1
	10	6	0	1	2	14

[b]B/W=black/white comparison; F/M=female/male comparison; H/W=Hispanic/white comparison; S=spring; F=fall

Pretest Reanalysis

In order to compare the results of the bias analyses across groups, the pretest data were reanalyzed using the calculated distance from the perpendicular line as the criterion rather than the judgment of outliers from the delta plots (the items falling outside the ellipse formed by the set). The reanalysis resulted in a greater number of items identified as biased than in the original analysis. In the original analysis, 12 items identified as biased were included in the final test form; in the pretest reanalysis, 43 items were identified as biased according to the criteria of a D value equal to or greater than 1.00. (Use of a criteria of D equal to or greater than 1.25 would have more closely approximated the number of items identified originally.) Inspection of the plots for the two tests that showed the greatest difference in items identified as biased (Reading Comprehension for the majority/Hispanic comparison and Science for the majority/black comparison) showed that the ellipse originally was drawn at a different slope than the principal axis. This resulted in the identification of not only more, but actually different, items than in the original analysis. It is interesting to note that the three items identified as biased in the male/female comparison were so identified in both analyses, whereas the degree of correspondence in the majority/black and majority/Hispanic samples was only 6 of 40. It is possible that the large sample sizes in all male and female analyses contributed to the similarity of results.

For item selection, anchor test scores were used to adjust data from individual samples so that the results could be treated as if all come from average-ability groups. The average anchor test score and the average number of students in the sample for each pretest unit by majority/minority group are as follows: average anchor test mean (average N in parenthesis) in the majority sample, 283.2 (382); in the black sample, 225.2 (99); in the Hispanic sample, 259.5 (72); in the male sample, 277.8 (193); and in the female sample, 289.7 (184). Because the sample sizes were small, it was not possible to select subsamples of comparable abilities. Therefore, an attempt was made to control for ability by adjusing p values to reflect an average ability level. The item data for the majority and black samples were adjusted according to the ratio of the national to the group anchor test scores. The national average used in this analysis was 295, the 1978 standardization average for grade 4. The reanalysis for the Social Studies test showed consistent results when the items identified as biased were compared using adjusted versus unadjusted item data. Using a D criteria of equal to or greater than 1.00, 23 of the 146 items were biased in both analyses; six were identified only when using adjusted data and one was identified only when using unadjusted data.

A comparable reanalysis was performed on the spring and fall analysis data. As in the pretest sample, the ability levels of the comparison groups within the spring and fall analyses were

not identical. The average raw scores on the Educational Ability Series test which was administered to the entire sample, and the number of students in each group are presented in Table 3 for both the spring and fall samples.

TABLE 3

Ability Averages for Spring and Fall Bias Samples

	Spring (Grade 4)		Fall (Grade 5)	
Sample	N	mean	N	mean
White	2312	30.2	1897	32.2
Black	406	20.3	418	23.7
Hispanic	116	25.1	78	26.4
Male	1487	28.5	1220	30.1
Female	1440	28.6	1279	30.6
National		28.6		30.4

Adjustment of spring and fall data for the same test using the EAS score as the anchor yielded less positive results for both sets of data. In the spring data, four items were biased only when adjusted and two were biased when adjusted and unadjusted; in the fall data, two items were biased only when adjusted.

Pretest-Reanalysis Comparisons

There were consistent patterns in the relationship of performance between the groups on each test across analyses. Table 4 presents the delta means for each group and the correlation between the item delta values for the pretest, fall, and spring reanalyses.

In general, the spring and fall delta averages for each group were similar. In most cases, the fall delta averages were lower than the spring averages, which indicates that the items were easier for the fall group (recall that delta has an inverse relationship to the p value). This is understandable because the fall group was tested later than the spring group--at the beginning of the following school year. The spring group in general performed better than the pretest group that had been tested the previous April. The spring and fall correlations of delta scores were, with a few exceptions, in the .90s; the pretest correlations ranged from .608 to .943 and reflect the fact that there were a great number of outliers in the pretest analysis.

TABLE 3

Delta Means across Comparison Groups and Samples

Test	Sample	Comparison Groups Black	White	r
Vocabulary	Pretest	14.201	12.183	.687
	Spring	13.033	10.934	.899
	Fall	13.435	11.262	.890
Comprehension	Pretest	13.898	11.878	.672
	Spring	12.742	11.065	.960
	Fall	13.358	11.291	.946
Math Concepts	Pretest	13.692	12.307	.831
	Spring	13.343	11.772	.943
	Fall	12.920	11.449	.945
Lang. Usage	Pretest	13.736	11.925	.642
	Spring	13.213	11.376	.834
	Fall	12.783	11.249	.867
Reference Materials	Pretest	14.199	12.281	.738
	Spring	13.060	11.263	.940
	Fall	12.535	10.969	.920
Social Studies	Pretest	14.420	12.809	.774
	Spring	13.716	11.969	.863
	Fall	13.370	11.724	.916
Science	Pretest	15.124	13.457	.570
	Spring	14.604	12.841	.906
	Fall	14.104	12.613	.911

Test	Sample	Comparison Groups Female	Male	r
Computation	Pretest	12.141	12.844	.943
	Spring	12.080	12.683	.988
	Fall	13.084	12.399	.983
Social Studies	Pretest	12.631	12.940	.894
	Spring	12.225	12.296	.968
	Fall	11.967	12.185	.951
Science	Pretest	13.509	13.508	.888
	Spring	13.262	13.005	.943
	Fall	12.978	12.885	.921

Test	Sample	Comparison Groups Hispanic	White	r
Vocabulary	Pretest	13.340	12.169	.608
	Spring	12.679	11.262	.923
	Fall	12.521	10.934	.872
Comprehension	Pretest	13.197	11.841	.694
	Spring	12.126	11.291	.945
	Fall	12.302	11.065	.887
Computation	Pretest	13.148	12.484	.868
	Spring	12.541	12.189	.941
	Fall	12.654	12.399	.927
Mechanics	Pretest	12.787	11.819	.647
	Spring	11.507	11.024	.943
	Fall	11.607	10.992	.905
Spelling	Pretest	13.346	12.625	.744
	Spring	11.889	11.329	.956
	Fall	12.057	11.199	.843
Reference Materials	Pretest	13.332	12.251	.777
	Spring	11.909	11.262	.962
	Fall	11.857	10.969	.909
Social Studies	Pretest	13.630	12.584	.762
	Spring	12.646	11.969	.954
	Fall	12.484	11.724	.914
Science	Pretest	14.248	13.418	.737
	Spring	13.513	12.841	.926
	Fall	13.506	12.613	.910

Consistency of Identification of Bias

Little overlap was found in the identification of biased items across analyses. With a criterion of distance (D) equal to or greater than a value of 1.00, only one item overlapped between both the pretest and spring reanalysis and the pretest and fall analysis, and three items overlapped between the spring and fall analyses. As in Scheuneman's study where 13 of the 51 items with high bias indices (greater than .75) in either administration were biased in both, there was less overlap than expected. Use of the same criterion that Scheuneman used improved the consistency of results in this study between spring and fall; 13 of the 32 items were biased in both analyses. The improvement in the overlap between pretest and spring/fall analyses was minimal; there was an overlap of 3 items between the pretest and spring analysis and four items between the pretest and fall analysis. In addition to the differences between the pretest and renalysis samples noted in the samples section, the fact that the pretest analysis contained substantially more items than the reanalyses also could have contributed to the lack of overlap. Since by the definition of bias assumed usign the delta method, "bias is discovered when an item does not fit the pattern established by others in the set," the differently constituted item sets could have placed the items in different contexts (Shepard, 1980). The items identified as biased in the majority/minority comparisons for the three analyses are presented in Table 5.

TABLE 5

Items Identified as Biased across Analyses [C]

Test/Comparison	Pretest	Spring	Fall
Vocabulary			
B/W	3 *7*	1 5M *14*	1M *5* *14*
H/W	23		*2* 9 *32*
Comprehension			
B/W	9 11 *33*	24	
H/W	8 10 11 *20* *21* *31* *35* *36* 49 50		*13* 17
Math Concepts			
B/W	18M 27	15M 18 29M	15 *29*
Computation			
F/M		29	
H/W	22 *37*		*6* *7* 16 22M 34
Mechanics			
H/W	16 17 22M 21 28		22
Usage			
B/W	18 30	*15* 18 *39*	15M 18 39M
Spelling			
H/W	*2* *13*	23M	23
Reference Materials			
B/W	*21*	9	9
H/W			9
Social Studies			
B/W	4 *19* 30	*5* 40	40M
F/M	*15*		15M
H/W	4 22		31
Science			
B/W	*3* 17 *20* *25* *32* *33* *35*	3M	3M
F/M		14	*10*
H/W	*8* *18* *27* *32*		*35*

[C] B/W=black/white comparison; F/M=female/male comparison; H/W=Hispanic/white comparison; numbers in italics indicate items biased against the minority group; numbers which are underlined indicate item which overlap across analyses; and numbers with an "M" suffix indicate marginally biased items (D values greater than .75 but less than 1.00)

Because of the small sample size in the Hispanic samples, less weight should be placed on the identification of items as biased against these samples. More than two-thirds of the total number of items identified as biased came from the white/Hispanic comparison; of these 37 items, 14 were biased in the fall analysis which contained less than 100 students in the Hispanic sample.

Direction of Bias

"Bias" (or unexpected performance difference, as Scheuneman chooses to call such a discrepancy in another 1980 paper, or item discrepancy, as Angoff (1980) refers to it), is not always found to favor the majority group. Inspection of Table 5 shows an almost equal split between items biased toward and against the minority group. As is shown in the pretest reanalysis results, for items used in the final test forms, balancing was accomplished in the test as a whole if not at the subtest level. In two subtests the items were biased in one direction only--the Language Arts Mechanics items were biased toward the Hispanic group and the Science items were biased against the black group.

Because some of the items selected for the final forms were not pretested with each minority group, pretest-reanalysis comparisons on only the final items were not possible. The D values for the pretest and spring and fall analyses for each item identified as biased in any of these analyses are summarized in

Table 6. The degree of consistency of D values is illustrated in the range of D values in the other analyses for items biased in one of the analyses. For example, 11 of the items identified as biased in the pretest analysis (D equal to or greater than 1.00) were in the D value range of .00 to .24 in the spring analysis. D values in this range would indicate only random variation between groups. Upon reanalysis, therefore, the items were not only not biased, but the variations could be considered random.

TABLE 6

Summary of Degree of Consistency of D Values

		Pretest		Spring		Fall	
		S[d]	F	P	F	P	S
opposite sign:	.75-.99	1	1	0	1	0	0
	.50-.74	0	0	0	0	1	0
	.25-.49	6	3	0	0	2	3
	.00-.24	11	10	1	0	2	2
same sign:	.00-.24	12	16	1	0	0	0
	.25-.49	9	7	2	2	1	8
	.50-.74	3	4	2	2	4	1
	.75-.99	0	1	1	4	2	4
	1.00+	1	1	1	3	1	3

[d]S=spring; F=fall; P=pretest; underlined numbers indicate items identified in baseline comparison.

Baseline Comparison

In response to suggestions for improvement of the delta method (Angoff, 1980), a baseline study was conducted in which comparisons of white-white (W/W), black-black (B/B), Hispanic-Hispanic (H/H), male-male (M/M), and female-female (F/F) data were made. This approach, which uses random samples within the same minority/majority group, is a variation of the recommendation for selecting samples that vary on a variable other than ethniticy, such as socioeconomic status or geographic region. If either sample definition is used, the outer limit of the index of bias for two samples within the same ethnic group is used as the outer limit of acceptability for the study of two ethnically different samples.

In the baseline comparison, no differences (D values equal to or greater than 1.00) were found in the W/W, B/B, M/M or F/F analyses; significant differences, however, were found in 20 items in the H/H comparisons. Six of the items identified as biased in this comparison were among those identified as biased in either the spring or fall analyses. Five (more than one-third of the items identified as biased) of these were from the fall sample. These results are incorporated in Table 6 and confirm the suspicion that there were peculiarities in the fall Hispanic sample, very likely the result of the small sample size.

Differences in Item Characteristics

Examination of the item characteristics for the majority/minority samples for each item identified as biased in either the pretest, the spring, or the fall analysis showed that approximately half of the items were more difficult for the minority group; that is, the more difficult items tended to discriminate against the minority group whereas the easier or as easy items tended to discriminate against the majority group. When the term "unexpected performance difference" is used in place of "bias", the result is logical, considering the performance differences between groups. In some cases, each distractor pulled as well as in the minority and majority group; in others, a specific distractor pulled differentially more in the minority group than in the majority group. An inspection of the chosen alternatives showed no consistent pattern in responding. Twice as many items were equally discriminating (point biserials) in the majority and minority groups than were less discriminating in the minority group; few were more discriminating in the minority group.

Regardless of the direction of bias, there was little difference between the white and black samples in terms of percent of omits. When items were biased against the Hispanic sample, there was no difference in the percent of omits, but when the items were biased toward the Hispanic sample, the white sample had approximately twice the percent of omits that the Hispanic sample had. When confronted with items to which they did not know the

correct response, students in the Hispanic sample, more often than students in the white or black samples, tended to omit the item.

In terms of "not reached" items, or test speededness, the black sample had approximately 2 to 3 times the percent of "not reached" items, almost twice as many when the items favored the black sample and almost three times as many when the items favored the white sample . In the Hispanic comparison, when the items favored the white sample, the percentages of "not reached" were approximately the same, but when the items favored the Hispanic sample, the white sample didn't reach approximately one-and-a-half as many items. Male versus female sample percents of either "not reached" or omits showed no difference. There was a greater tendency among students in the black sample as compared to the white sample and, similarly, for the white sample as compared to the Hispanic sample, to not reach the "biased" item.

Skill Areas Measured

The skill areas measured by items identified as biased were also inspected. A definite pattern was found in only one subtest--the five Language Arts Mechanics items that were biased toward the Hispanic sample all measured punctuation. The summary of skill areas for biased items are presented in Table 7.

The following skill areas were identified as representing a moderate percentage of items identified as biased by any analysis: 7 of 15 Vocabulary items measuring synonyms; 5 of 11 Reading Comprehension items measuring detail, 4 of 15 items measuring perceiving relationships, and 3 of 13 items measuring drawing conclusions; 3 of 8 Math Concepts items dealing with geometry; 4 of 22 Math Computation items dealing with whole numbers; 5 of 15 Language Arts Mechanics items dealing with punctuation; 3 of 17 Social Studies items dealing with geography/economics, 3 of 16 items dealing with sociology/anthropology, and 2 of 7 items dealing with history/government; 3 of 11 Science items dealing with scientific knowledge, 2 of 13 items dealing with process, and 5 of 16 items dealing with concepts and principles.

TABLE 7

Skill Areas Measured by Biased Items[e]

Test/Comparison	Direction of Bias: Against Minority	Against Majority
Vocabulary		
B/W	synonyms (3)	synonyms (2)
H/W	synonyms (1)	synonyms (1)
	idioms (1)	figurative language (1)
Comprehension		
B/W	detail (1)	detail (1)
		author's purpose (1)
		summarizing (1)
H/W	detail (2)	detail (1)
	cause & effect (2)	cause & effect (1)
	inference (1)	inference (1)
	compare/contrast (1)	summarizing (1)
		author's purpose (1)
		how character felt (1)
Math Concepts		
B/W	geometry (1)	geometry (2)
		fractions (1)
Computation		
F/M		decimals/percents (1)
H/W	whole numbers (2)	whole numbers (2)
	mixed numbers (1)	decimals/percents (1)
Problem Solving		
Mechanics		
H/W		punctuation (5)
Usage		
B/W		pronouns/modifiers (1)
		verbs (1)
		structure (1)
		transformation (1)
Spelling		
H/W	homophones (1)	variant pronunciation (1)
	graphemes (1)	
Reference Materials		
B/W	index (1)	pronuncaition key (1)
H/W		pronunciation key (1)
Social Studies		
B/W	geography/economics (2)	geography/econ. (1)
		sociology/anthrop. (1)

		history/government (1)
F/M	sociology/anthrop. (1)	
H/W		sociology/anthrop. (1)
		history/government (1)
Science		
B/W	process (2)	concepts/principles (1)
	knowledge (1)	
F/M	knowledge (1)	concepts/principles (1)
H/W	knowledge (1)	
	concepts/principles (1)	
	process (3)	

[e]B/W=Black/white comparison; F/M=female/male comparison; H/W=Hispanic/white comparison; numbers in parentheses indicate number of biased items within the skill area.

Conclusions and Recommendations

This study supported the conclusions of Scheuneman's investigation of consistency across test administrations with varying sample sizes. The results of both studies indicate that the degree of consistency was less than expected. Sample size, in this study, was shown to be a factor in that the most consistent results were from samples of more than 1000 students and the least consistent results were from samples of less than 100 students. A couple of the recommendations for improvement of the delta method were applied with nonconclusive results. Adjusting for ability levels resulted in the identification of essentially the same items but did not improve the consistency across test administrations; making baseline comparisons (majority/majority and minority/minority) produced significant differences only in the minority sample with the small sample size.

Because of the wide use of the delta method, it would seem important to conduct a study in which all of Angoff's recommendations were implemented--controlling for ability, making baseline comparisons, using an external variable to control for ability, looking at variations within an ethnic group, and replicating the study to determine the reliability of the index. In practice, it may only be possible to study majority/majority effects to any extent because of the difficulty of obtaining large minority samples. Controlling for ability using the method

employed in this study is an alternative to the more difficult method of pulling matched samples on an external measure. Use of an anchor test or set of common items which are administered to at least two samples (either random or varying on another variable, such as SES or geographic area) from the majority group and one sample from each minority groups could address the crucial aspects of the recommendations.

References

Angoff, W.H. The use of difficulty and discrimination indices in the identification of biased test items. Paper presented at the Johns Hopkins University Symposium on Educational Research, "Test Item Bias Methodology: State of the Art", Washington D C, November 7, 1980.

Angoff, W. H. and Ford, S.F. Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10, 95-105.

Ironson, G.H. and Subkoviak, M. A comparison of several methods of assessing item bias. Journal of Educational Measurement, 1979, 16, 209-225.

Merz, W.R. and Grossen, N.E. An empirical investigation of six methods for examining test item bias. Report submitted to the National Institute of Education, Grant NIE-78-0067, California State University, Sacramento, California, 1979.

Rudner, L.M. and Convey, J.J. An evaluation of select approaches for biased test identification. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.

Rudner, L.M., Getson, P.R. and Knight, D.L. A Monte Carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, 1980, 17, 1-10.

Scheunemann, J.D.Consistency across administration of certain indices of bias in test items. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

Scheunemann, J.D.A posteriori analysis of biased items. Paper presented at the Johns Hopkins University Symposium on Educational Research, "Test Item Bias Methodology: State of the Art", Washington, D C, November 7, 1980.

Shepard, L.A., Camilli, G., and Averill, M. Comparison of six procedures for detecting item bias using both internal and external ability criteria. Paper presented at the annual meeting of the National Council for Measurement in Education, Boston, April 1980.

Shepard, L.A. Definition of bias. Paper presented at the Johns Hopkins University Symposium on Educational Research, "Test Item Bias Methodology: State of the Art", Washington, D C, November 7, 1980.